Textdatenbanken

Sommersemester 2020

- Fragen zur Vorlesung -

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

FRAGE: Richtig oder falsch?

In einem Satz bezeichnet man die von Leerzeichen begrenzten Zeichenketten als Wörter.

Antwort: Falsch

Was kommt im Satz außer Wörtern und Leerzeichen noch vor?

Zusätzlich müssen die Satzzeichen abgetrennt werden, die ohne Leerzeichen neben einem Wort stehen.

FRAGE: Richtig oder falsch?

Nur Punkt, Ausrufezeichen oder Fragezeichen können am Ende eines deutschen Satzes stehen.

Antwort: Falsch

Kann auf Punkt, Ausrufezeichen oder Fragezeichen noch ein weiteres Zeichen folgen?

Zusätzlich müssen (alle Arten von) An- und Ausführungszeichen beachtet werden.

FRAGE: Richtig oder falsch?

Wenn ich beim Lesen eines Textes zu einem Punkt gelange, dann kann ich immer entscheiden, ob ein Satzende vorliegt, ohne weiterlesen zu müssen.

Antwort: Falsch

Können Abkürzungen am Satzende stehen?

In einigen Fällen kann erst durch Weiterlesen entschieden werden, ob ein Satzende vorlag.

FRAGE: Richtig oder falsch?

Für die Abkürzungsliste zur Zerlegung in Sätze sind Abkürzungen wie *NATO* und *kg* unwichtig.

Antwort: Richtig

Versuchen Sie, einen Satz zu bilden, der *NATO* oder *kg* im Inneren enthält und bei dem diese Wörter fälschlicherweise für ein Satzende gehalten werden könnten.

Nur Abkürzungen, die auf einen Punkt enden, können mit Satzenden verwechselt werden.

FRAGE: Richtig oder falsch?

Die Anzahl der Einträge in der Wortliste ist gleich der Gesamtlänge aller Sätze, gemessen in Wörtern.

Antwort: Falsch

Stehen Wörter mehrfach in der Wortliste?

Wörter, die mehrfach in Sätzen vorkommen, stehen nur einmal in der Wortliste.

FRAGE: Richtig oder falsch?

In der Textdatenbank gibt es für ein Wort möglicherweise weniger Beispielsätze, als die Anzahl des Wortes angibt.

Antwort: Richtig

Kommt ein Wort immer nur in verschiedenen Sätzen vor?

Kommt ein Wort beispielsweise insgesamt zweimal vor, und zwar im selben Satz, so gibt es trotz Anzahl 2 nur einen Beispielsatz.

FRAGE: Richtig oder falsch?

- Die Anzahl der Wörter im einer Textdatenbank ist ungefähr gleich der Anzahl der Leerzeichen in den Sätzen.
- Die Anzahl der Wörter im einer Textdatenbank ist ungefähr 10% größer als die Anzahl der Leerzeichen.

FRAGE

Vergleichen Sie:

Archiv – Korpus – Textdatenbank!

FRAGE

Nennen Sie vier konkrete linguistisch interessante Anfragen, die sich nicht ohne zusätzliche Hilfe mit Google lösen lassen.

Frage

Welche Fragestellungen lassen sich besser korpusbasiert als von Experten beantworten?

FRAGE: Richtig oder falsch?

Das Dateiformat ist für die Aufnahme eines Textes in eine Textdatenbank völlig unwichtig.

Antwort: Falsch

Wie erzeugen Sie Klartext aus verschiedenen Textverarbeitungsformaten?

Die Konvertierung macht für unterschiedliche Ausgangsformate verschieden große Schwierigkeiten. Deshalb kann auch die Qualität des entstehenden Klartextes vom Format abhängen.

FRAGE

Warum sind große Korpora besonders nützlich?

Frage

Wie viele Wörter in einer nach Häufigkeit sortierten deutschen Wortliste benötigt man, um alle Buchstaben des Alphabets vorzufinden?

a) 100
b) 1.000
c) 10.000
d) 100.000

Beispiele

- X auf Rang 940: Experten (1143: Praxis, 1208: Text)
- Y auf Rang 635: Bayern (689: System, 1395: Analysten)
- Q auf Rang 2816: Konsequenzen (3953: konsequent, 4347: quasi)
- É auf Rang 5691: Café (7338: André, 7853: René)

FRAGE

Mittels Levenshtein-Ähnlichkeit lassen sich Präfixe und Suffixe ermitteln.

FRAGE

Nennen Sie unterschiedliche Textgenres!

FRAGE

Wodurch unterscheiden sich:

- Zeitungstext
- Romane
- Gedichte
- Fachtext
- Filmuntertitel

FRAGE

Unterscheiden sich Textgenres durch:

- ihre Stoppwörter?

FRAGE

Unterscheiden sich Textgenres durch:

- ihre häufigsten Substantive?
- welche anderen Wortarten?

FRAGE

Für welche Textgenres lassen sich große Korpora in mehreren Sprachen erzeugen?

In wie vielen Sprachen etwa?

Frage

Die Arbeitsschritte zur Erstellung von Textdatenbanken sind immer gleich. Alle Sprachen können genau wie die deutsche Sprache behandelt werden.

Antwort: NEIN

- Andere Zeichensätze
- Keine Unterscheidung von Groß- / Kleinschreibung
- Andere Satzzeichen
- Andere oder keine Worttrennung

FRAGE: Richtig oder falsch?

In vielen Sprachen findet man zum Stichwort *IBM* unter den Satzkookkurrenzen weitere Computerfirmen.

Antwort: Richtig

An welche anderen Firmen denken Sie bei *IBM*?

Unabhängig von der Sprache wird die Firma *IBM* stets mit den gleichen Firmen genannt, beispielsweise *Sun*, *Dell*, *Microsoft*.

FRAGE: Richtig oder falsch?

Seltene Abkürzungen sind leichter aufzulösen als sehr häufige.

Antwort: Richtig

Wie wird dafür gesorgt, dass der Leser eine seltene Abkürzung versteht?

Weil seltene Abkürzungen häufig unbekannt sind, werden sie sicherheitshalber noch in der Langform angegeben, z.B. »*Slowakische Nationalpartei (SNS)*«. Häufige Abkürzungen wie *BRD* werden hingegen als bekannt vorausgesetzt.

FRAGE: Richtig oder falsch?

POS-Tagging macht es leicht, nach Personennamen zu suchen.

Antwort: Richtig

Welches Tag bekommen Vor- und Nachnamen?

Vor- und Nachnamen tragen das Tag NE, vor dem Namen steht häufig ein Titel oder eine Berufsbezeichnung. Außerdem stehen für Vor- und Nachnamen sowie Titel und Berufe umfangreiche Listen zur Verfügung.

FRAGE: Richtig oder falsch?

Mittels POS-Tagging können alle Personennamen in einem Text gefunden werden.

Antwort: Falsch

Vergleichen Sie die Sätze „*Ich kenne Wolfgang Rindfleisch*.“ und „*Heute isst Manfred Rindfleisch*.“

Auch beim POS-Tagging werden Fehler gemacht. Entweder weil Wörter wie *Rindfleisch* und *Vogel* nur selten Nachname sind oder weil Wörter dem Tagger völlig unbekannt sind.

FRAGE

Wozu kann man die Beispielsätze zu einem Wort nutzen?

FRAGE

Stört Datenmüll im Korpus?

Welche Sorten Müll kennen Sie?

FRAGE

Wie findet man doppelte Sätze in einem Korpus?

Wie findet man Quasi-Dubletten (d.h. sehr ähnliche Sätze)?

FRAGE

Nennen Sie Gründe, warum Satzdubletten und Quasidubletten aus einer Textdatenbank entfernt werden sollten!

FRAGE

Wie gut ist die folgende Regel?

> Wenn X typisches Werkzeug für die Tätigkeit V ist und Y Kohyponym von X, dann ist auch Y typisches Werkzeug für die Tätigkeit V.

Beispiel:

V: schneiden

X: Messer

Y: Schere

FRAGE

Wofür lassen sich Frequenzangaben zu Wörtern benutzen?

FRAGE

Welche Kriterien sorgen für die Auffälligkeit eines Wortes an einem Tag, verglichen mit anderen Tagen?

FRAGE

In welchem Verhältnis stehen das Jahr der Veröffentlichung von Zeitungstexten und die im Korpus erwähnten Jahreszahlen?

FRAGE

Nennen Sie technische Probleme beim Messen linguistischer Masszahlen!

FRAGE

Wofür sind die folgenden Repräsentationen der Messergebnisse nützlich?

- Optisch: Graphische Darstellung
- Minimal: Beschreibung durch möglichst wenige Parameter
- Exemplarisch: Angabe von Beispielen in einer Tabelle

FRAGE

Welche Annotationen können in einem Korpus vorgenommen werden?

FRAGE

Welche Eigenschaften hat ein typischer Satzanfang?

Welche Eigenschaften hat ein typisches Satzende?

FRAGE

Sind die Relationen Synonymie und Kohyponymie transitiv?

FRAGE

Warum ist Kohyponymie die häufigste Relation im Thesaurus?

FRAGE

Nennen Sie Quellen für Paralleltext!

Nennen Sie Quellen für quasiparallele Texte!

FRAGE

Was verstehen Sie unter Ankern bei Paralleltext?

Wozu können sie benutzt werden?

FRAGE

Nennen Sie Vor- und Nachteile der Bibel als Paralleltext!